Nucleic Acids Research, 2004, Vol. 32, Web Server issue ***W449–W456***
DOI: 10.1093/nar/gkh409

Onto-Tools: an ensemble of web-accessible, ontology-based tools for the functional design and interpretation of high-throughput gene expression experiments

Purvesh Khatri, Pratik Bhavsar, Gagandeep Bawa and Sorin Draghici*

Department of Computer Science, Wayne State University, 431 State Hall, Detroit, MI 48202, USA

Received February 6, 2004; Revised and Accepted March 30, 2004

ABSTRACT

The Onto-Tools suite is composed of an annotation database and five seamlessly integrated web-accessible data mining tools: Onto-Express (OE), Onto-Compare (OC), Onto-Design (OD), Onto-Translate (OT) and Onto-Miner (OM). OM is a new tool that provides a unified access point and an application programming interface for most annotations available. Our database has been enhanced with more than 120 new commercial microarrays and annotations for *Rattus norvegicus*, *Drosophila melanogaster* and *Carnorhabditis elegans*. The Onto-Tools have been redesigned to provide better biological insight, improved performance and user convenience. The new features implemented in OE include support for gene names, LocusLink IDs and Gene Ontology (GO) IDs, ability to specify fold changes for the input genes, links to the KEGG pathway database and detailed output files. OC allows comparisons of the functional bias of more than 170 commercial microarrays. The latest version of OD allows the user to specify keywords if the exact GO term is not known as well as providing more details than the previous version. OE, OC and OD now have an integrated GO browser that allows the user to customize the level of abstraction for each GO category. The Onto-Tools are available online at http://vortex.cs.wayne.edu/Projects.html.

INTRODUCTION

In molecular biology and genetics, our data gathering capabilities have greatly surpassed the available data analysis techniques. Examples of modern high-throughput techniques able to produce data at a phenomenal rate include shotgun sequencing (1,2) and gene expression microarrays (3–9). The continuous use of these high-throughput data collection techniques over the years has produced a large amount of heterogeneous data. The challenge faced by today's researchers is to develop effective ways to analyze the vast amount of data that has been and will continue to be collected (10–12).

Onto-Tools is an open-access software suite that partially addresses this problem. This is achieved by using a probabilistic functional analysis that bridges the gap between low-level, high-throughput gene expression data and high-level functional knowledge. The Onto-Tools suite includes (i) Onto-Express (OE), which can be used to translate lists of differentially regulated genes into a better understanding of the underlying biological phenomena through the use of Gene Ontology (GO); (ii) Onto-Design (OD), used to select the best set of genes to be included on a custom microarray designed for the study of a given biological phenomenon; (iii) Onto-Compare (OC), used to analyze the functional bias of various focused commercial microarrays and select the one that is most appropriate for a given biological hypothesis; (iv) Onto-Translate (OT), which is used to translate lists of genes from one reference system to another (e.g. from GenBank accession numbers to UniGene cluster IDs to Affymetrix probe IDs) and (v) Onto-Miner (OM), which provides a unified access point and an application programming interface (API) allowing queries for various information such as the gene name, official symbol, reference accession number and coded protein. These tools are freely available at http://vortex.cs.wayne.edu/Projects.html. The suite currently supports *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster* and *Caenorhabditis elegans*. It has more than 1,100 registered users from over 50 countries, and it processes an average of over 100 data sets daily. Previous publications have described in detail the motivation, implementation and validation of these tools (13–16). This paper first briefly reviews these tools for the benefit of new

*To whom correspondence should be addressed. Tel: +1 313 577 5484; Fax: +1 313 577 6868; Email: sod@cs.wayne.edu

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated.

© 2004, the authors

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 29, 2015

users and then focuses on a number of recent additions and enhancements that expand the capabilities of these tools.

OVERVIEW

Onto-Express

Independent of the platform and the analysis methods used, the result of a microarray experiment is, in most cases, a list of genes found to be differentially expressed between two or more conditions under study. The challenge faced by the researcher is to translate this list of differentially regulated genes into a better understanding of the underlying biological phenomena. The translation from a list of differentially expressed genes to a functional profile able to offer insight into the cellular mechanisms is a very tedious task if performed manually. Typically, one would take each regulated gene, search various public databases and compile a list of, for instance, the biological processes that the gene is involved in. In order to construct a master list of all the biological processes in which at least one gene is involved, this task must be performed repeatedly for each gene. Further processing of this list provides a list of those biological processes that are common between several of the regulated genes. It is expected that those biological processes that occur more frequently in this list will be more relevant to the studied condition. For instance, if all the genes found to be regulated were involved in apoptosis, one would conclude that the condition studied has significant impact on the apoptotic pathway. The same type of analysis is necessary for other functional categories such as molecular function and cellular component.

We designed OE as a tool implementing a rigorous approach to this process (13, 14, 16). This is accomplished by mining known data and compiling a functional profile of the studied condition. Many months of tedious and inexact manual searches are replaced by a few minutes of fully automated analysis. OE's input is a list of genes found to be regulated in a specific condition. The result of this analysis is a functional profile of the condition studied. The results are provided in graphical form or saved as a file in semicolon-delimited format that can be imported into Excel. OE constructs a profile for each of the GO categories (17): cellular component, biological process and molecular function. As biological processes can be regulated within a local chromosomal region (e.g. imprinting), an additional profile is constructed for the chromosome location. These functional profiles are accompanied by the computation of significance values for each functional category. Such values allow the user to distinguish between significant biological processes and random events. OE uses a database designed, implemented and maintained in our laboratory (http://vortex.cs.wayne.edu:8080). We use data from GenBank, UniGene, LocusLink and PubMed. We strive to keep our database up to date with the latest information and we update its content every time one of our sources releases a new version of their data.

Onto-Compare

Many microarray users embark upon 'hypotheses generating experiments' in which the goal is to find subsets of genes differentially regulated in a given condition. However, another major application of this type of data mining is in experiment design. An alternative to the 'hypotheses generating experiment' is the 'hypothesis driven experiment' in which one first constructs a hypothesis about the phenomenon under study and then performs directed experiments to test the hypothesis. However, specific hypotheses and a small number of pathways may still involve hundreds of genes. This is still too many for RT–PCRs, western blotting and other gene-specific techniques, so the microarray technology may still be the preferred approach.

Currently, no two arrays offer exactly the same set of genes. When a hypothesis about a certain mechanism does exist, we argue that one should use the array(s) that best represent the corresponding pathways. This can be accomplished by analyzing the list of genes on all existing arrays and providing information about the pathways and biological mechanisms covered by the genes on each microarray. If array A contains 10 000 genes but only 80 are related to a given pathway and array B contains only 400 genes but 200 of them are related to the pathway of interest, the experiment may provide more information if performed with chip B instead of chip A. This can also translate into significant cost savings.

Many commercial microarray manufacturers have realized the need for such focused arrays and have started to offer many such arrays. For instance, ClonTech currently sells focused human microarrays for the investigation of the cardiovascular system, cell cycle, cell interaction, cytokines/receptors, hematology, neurobiology, oncogenes, stress, toxicology, tumors and so on. Many other companies have picked up on the same trend and offer focused arrays, e.g. Perkin-Elmer, Takara Bio, SuperArray Inc., Sigma Genosys. Typically, a focused array includes a few hundreds of genes covering the biological mechanism(s) being studied. However, two microarrays produced by different companies are extremely unlikely to use the same set of genes. In consequence, various pathways are represented to various degrees on different arrays even if the arrays are all designed to investigate the same biological mechanisms. This is an unavoidable **functional bias**. Such a bias will be associated with each and every array that includes less than the full genome of a given organism.

Onto-Compare helps researchers assess the biological bias of various commercial arrays using the Onto-Tools database as a back-end. In addition to the data collected from various online databases, the Onto-Tools database is also populated with the lists of genes for each microarray (GenBank accession numbers, UniGene cluster IDs and LocusLink IDs) as provided by their respective manufacturers. We support all commercial microarrays currently available (172 microarrays from 8 manufacturers) and we will add more as they become available. Each locus in the LocusLink database is annotated using ontology from the GO Consortium. The GO Consortium provides ontology for biological processes, molecular functions and cellular components. The data from these databases and the gene lists corresponding to all commercial microarrays have been parsed and entered into our database. After creating a list of locus identifiers for each array, the list has been used to generate the following profiles: biological process, cellular component and molecular function. The profiles for each microarray are precalculated and stored in the database. The list of genes deposited on a microarray is static, but the annotations for those genes keep changing and are updated

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 29, 2015

automatically as more information becomes available. Onto-Compare is currently the only tool providing this type of analysis.

Onto-Design

In many cases, researchers prefer to print their own arrays. One of the reasons for opting to print one's own custom array is that given the complexity of the biological research one may feel that none of the commercially available microarrays represents the targeted pathways and biological processes to the extent needed. Other reasons may be related to the dramatically reduced price of an in-house solution compared with commercial arrays and the ability to adapt the arrays to one's own experimental design and use of controls. In order to design a microarray that constitutes a powerful and effective interrogation tool, a researcher has to choose genes that are representative of the key mechanisms and pathways. At present, the choice of genes to include on a certain microarray is a very laborious process requiring a high level of expertise. Furthermore, this process is very time consuming, even for experts, since they have to consult many on-line databases as well as perform an extensive literature review in order to find the set of genes that are involved in specific biological processes of interest. Onto-Design is a tool that has been developed to assist in this gene selection process.

The OD interface allows the user either to upload a set of functional categories of interest (such as biological processes) or to browse through a graphical representation of a tree representing the GO hierarchy. [Actually the GO hierarchy is a directed acyclic graph (DAG), not a tree. The internal structure of the database represents the GO hierarchy correctly, but the interface is more conveniently represented as a tree. Categories linked through DAG links not apparent in the tree are automatically traveled by the system in the appropriate way.] OD returns a list of all known genes annotated for each GO term in the input. The user can perform various set operations in order to tune the microarray design. The usefulness of this tool extends beyond microarrays to any of several high-throughput assays available today, at either mRNA or protein level [e.g. PowerBlots (18)].

Onto-Translate

In the annotation world, the same piece of information can be stored and viewed differently across different databases. For instance, more than one Affymetrix probe ID can refer to the same GenBank sequence (accession number) and more than one nucleotide sequence from GenBank can be grouped in a single UniGene cluster. The result of OE depends on whether the input list contains Affymetrix probe IDs, GenBank accession numbers or UniGene cluster IDs. In order to illustrate this, let us consider an input specified as a list of 10 Affymetrix probe IDs. Let us assume that the results show that 4 out of the 10 probes are involved in biological process A and the remaining 6 probes are involved in biological process B. Therefore, the frequency of biological process A will be 4 and that of process B will be 6. In order to interpret this, a researcher might need to use the data sheet provided with each Affymetrix array (or the NetAffy website) to map these probe IDs onto accession numbers. This may reveal that the 4 probe IDs for process A correspond to only 2 different accession numbers and the 6 probe IDs for process B correspond to another 2, different accession numbers. Repeating the OE analysis using accession numbers will show that the frequency of both processes A and B is 2. Furthermore, mapping the accession numbers onto UniGene cluster IDs may show that the 2 accession numbers for biological process A come from 2 UniGene clusters, whereas the accession numbers for biological process B come from the same UniGene cluster. Repeating the OE analysis using cluster IDs will show the frequency of A as 2 and B as 1. Since these frequencies are used to calculate the statistical significance of the occurrence of a specific category, it is clear that the type of identifiers used can dramatically influence the conclusions.

This example illustrates that the user has to be aware of these relationships between the different forms of the data in order to correctly interpret the results. Furthermore, even if a user is aware of the relationships and knows how to convert them, most existing tools allow conversions only of individual genes. This makes the process of translating hundreds of genes absolutely unfeasible. Onto-Translate is a tool that allows the user to easily perform such large-scale translations. A user can input a list of genes specified by either Affymetrix probe IDs, GenBank accession numbers or UniGene cluster IDs, indicate the type of the list by clicking the appropriate radio button and request the translation on the input list in either of the two remaining forms by selecting the appropriate radio button for the output list.

Onto-Miner

Onto-Miner is a new addition to the ensemble of Onto-Tools. Usually, the annotations for genes are divided across several public databases. For example, the GenBank database provides the nucleotide sequence and literature citations for a nucleotide sequence. However, the gene from which the sequence is derived, its location on the chromosome if known, various tissues the gene is expressed in, its similarity to other organisms and other similar sequences in the database are all stored in UniGene. Furthermore, for the official name, functional annotations and reference sequence one has to access LocusLink. Finally, in order to find out the metabolic and signaling pathways that the gene participates in one has to search pathway databases such as KEGG (19–21), BioCarta (22) and BioCyC (23).

OM is a database that integrates all known information about a gene in a unique resource. At present, it contains information from UniGene, GenBank, dbEST, LocusLink, RefSeq, KEGG and GO. The associated tool supports querying the OM database for *Bos taurus*, *C.elegans*, *Danio rerio*, *D.melanogaster*, *H.sapiens*, *M.musculus* and *R.norvegicus*. For each organism, the user can search by clone ID, GenBank accession number, UniGene cluster ID, gene symbol, gene name or LocusLink ID. The input to OM can be a list of IDs as a text file with one ID per line or the user can copy and paste the list from any other application into the text area provided on the OM input interface. For each gene, OM provides its UniGene cluster ID, UniGene name, official gene symbol, GI ID of the representative protein for the gene, RefSeq accession number, LocusLink ID, chromosome location and annotations using GO. OM will soon be expanded to integrate various pathway databases such as KEGG and BioCarta and protein databases such as UniProt (24). An

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 29, 2015

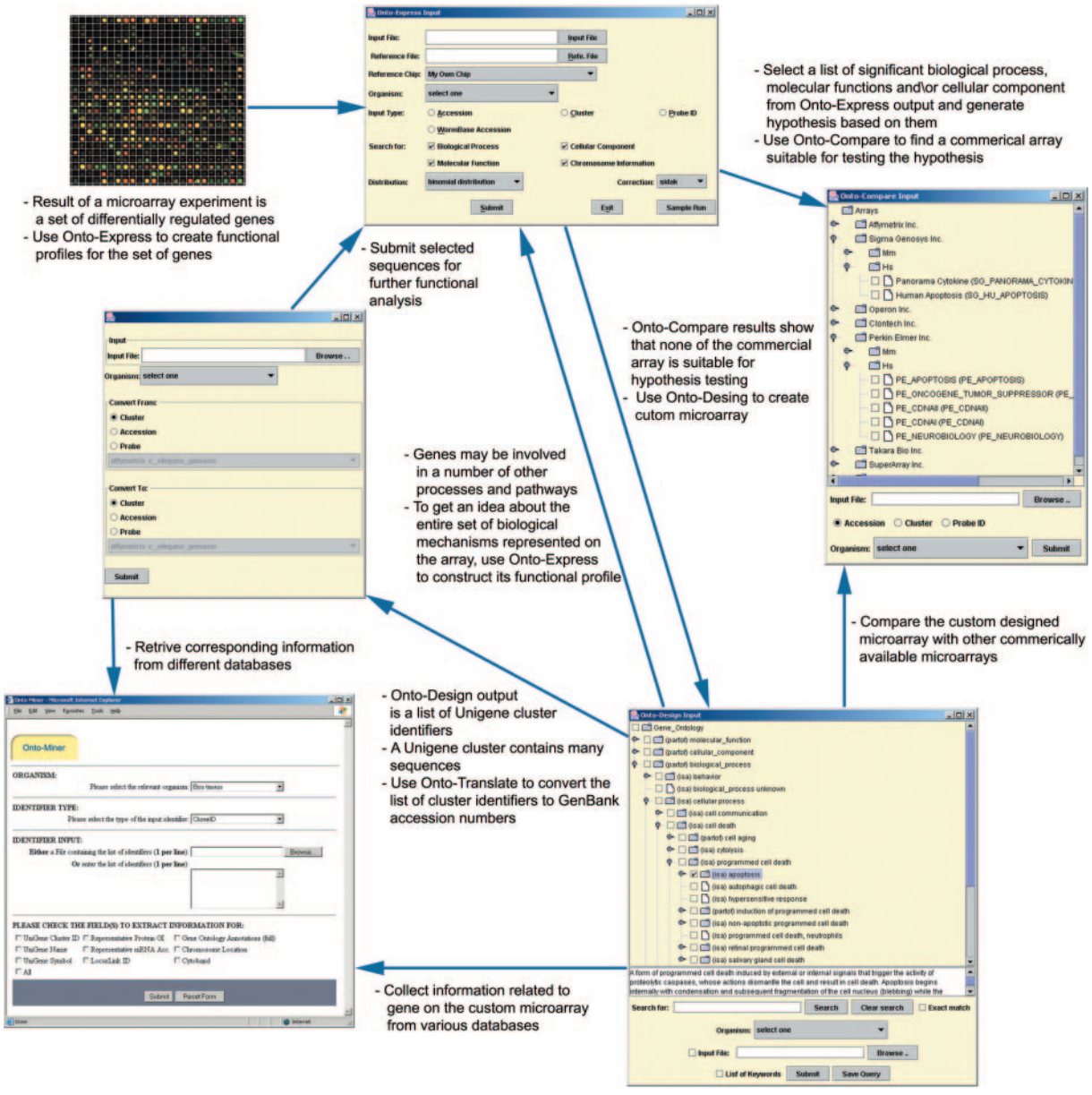

Figure 1. Onto-Tools integration (clockwise from top left). The user can analyze a set of differentially regulated genes using Onto-Express to find significantly affected biological processes. Onto-Compare can be used to find a suitable microarray for studying the hypotheses formulated based on these biological processes. If a suitable array is not found, Onto-Design can be used to design a custom array suitable for studying the hypotheses. One can refine the custom array design by creating the functional profile of the custom array and comparing it with the existing arrays. Once a suitable array is designed the user can either use Onto-Translate to obtain the list of GenBank accession numbers for the custom array or use Onto-Miner to obtain more details about each of the genes on the custom array such as the gene name, RefSeq accession number and chromosome location.

important feature of OM is the availability of an API that allows any third-party application to query our database. Thus, other researchers can write their own programs that will have access to all known gene annotations from a single, convenient source.

Tool integration

Each tool in the Onto-Tools ensemble addresses a specific problem. Although each tool is useful on its own, we tried to achieve a synergy by integrating them. Figure 1 provides a quick overview of some possible navigation between the tools.

The integration of Onto-Tools and its usefulness is best explained by an example. Consider a typical scenario where one uses a general-purpose comprehensive array such as Affymetrix HGU133 to screen a large number of genes in order to investigate a given condition and selects a set of differentially regulated genes. In order to understand the underlying biological phenomenon, the set of genes can be analyzed with OE, which will identify the biological processes and the molecular functions that are relevant and highly significant in the given condition (14). One can select these relevant and significant biological processes and molecular functions and submit them to OC with a single click. This will compare all

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 29, 2015

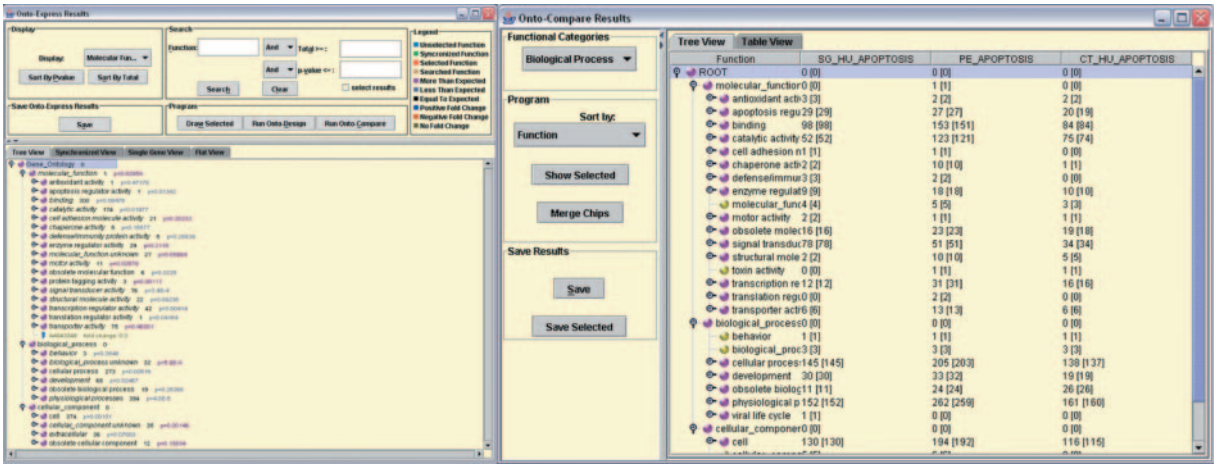

Figure 2. Integrated GO browser in Onto-Express (left) and Onto-Compare (right). The tree view in OE and OD allows the user to choose the desired level of abstraction for each GO category. The *P*-values for each term in OE are adjusted as the user chooses the desired level of abstraction for the term. The numbers of genes in each GO term for each array are modified appropriately as the terms are collapsed or expanded.

existing commercial microarrays which contain genes involved in these biological processes and molecular functions in order to find the array(s) that might be useful for probing the selected biological processes further. If none of the commercially available arrays covers the necessary pathway to a satisfactory degree, one can seamlessly pass the list of selected biological processes from OE to OD and design a custom array that is most suitable for the condition under study. After designing a custom array with OD, one can submit the custom array to OC and compare it with any of the existing commercial arrays to verify if the custom array is indeed better. However, the genes on the custom array may be involved in a number of other processes and pathways. In order to get an idea about the entire set of biological mechanisms represented on the custom array, one can submit the list of genes for the custom array to OE and create complete functional profiles of the custom array. Once the design of the custom array is satisfactory, one can use OT in order to translate the list of UniGene cluster IDs to GenBank accession numbers. Alternatively, one can also use OM to find out UniGene cluster ID, RefSeq accession number, protein GI ID, gene name and so on for each gene on the custom array.

ENHANCEMENTS AND NEW FEATURES

The back-end annotation database

Addition of new commercial microarrays. Over the past year, the Onto-Tools database has been enhanced with the addition of over 100 commercial microarrays. The database now contains gene lists and annotation data for 172 commercial microarrays. The database has also been updated with the latest gene lists from various microarray manufacturers including Affymetrix, Perkin-Elmer, Sigma-Genosys, Clontech Biosciences and SuperArray. The database is periodically updated as the public data repositories GenBank, dbEST, UniGene, LocusLink and GO are updated.

Support for more organisms. In addition to the new microarrays, annotation data for three more organisms have been added to the Onto-Tools database. The database now contains annotation data for *H.sapiens*, *M.musculus*, *R.norvegicus*, *D.melanogaster* and *C.elegans*. *Saccharomyces cerevisiae*, *Arabidopsis thaliana*, *Oryza sativa*, *Danio rerio*, *Dictyostellium discoideum*, *Geobacter sulfurreducens PCA*, *Pseudomonas syringae DC3000*, *Bacillus anthracis Ames*, *Coxiella burnetii RSA 493*, *Shewanella oneidensis MR-1*, *Vibrio cholerae*, *Leishmania major*, *Plasmodium falciparum*, *Schizosaccharomyces pombe*, *Trypanosoma brucei* and *Glossina morsitans* will be added in the near future.

Onto-Express

Integrated GO browser and customized abstraction levels. A major enhancement in the new release of OE was the integration of a GO browser in May 2003 (see Figure 2). The GO has a hierarchical structure where genes are annotated at various levels of abstraction. For instance, 'induction of apoptosis by hormones' is a type of 'induction of apoptosis', which in turn is part of 'apoptosis.' Apoptosis represents a higher level of abstraction, more general, whereas induction of apoptosis by hormones represents a lower level of abstraction, more specific. When annotating genes with GO terms, efforts are made to annotate the genes with the highest level of detail (lowest level of abstraction) possible. For example, if a gene is known to induce apoptosis in response to hormones, it will be annotated with the term 'induction of apoptosis by hormones' and not merely with one of the higher-level terms such as 'induction of apoptosis' or 'apoptosis'. In the previous release, OE results considered only the lowest level of abstraction with the highest level of detail. As a result, hundreds of GO terms were often included in the result, which made the interpretation of results rather difficult. For instance, if one tried to question whether apoptosis, as a global process, is significantly impacted, one was forced to scan the entire list of results looking for anything involved in apoptosis. Furthermore, since in many cases these extremely specific categories are represented by only one or two genes, often they do not appear as statistically significant. In order to answer the question posed, one needs to consider a higher level of abstraction (e.g. 'apoptosis') and calculate the *P*-value for this level.

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 29, 2015

Integration of the GO browser into OE allows the users to customize the level of abstraction for the given biological hypothesis. Certain branches can be expanded to provide maximum details and distinguish between specific subcategories, while others may be kept more general. OE dynamically calculates the new *P*-values for each term as the user chooses the desired level of abstraction for each category. The GO browser is displayed as a 'Tree view' in the OE graphical user interface (GUI). At any moment in time, the user has two types of results available. The flat view contains the results of the analysis (including significance values) at the lowest level of abstraction (most specific annotations). The synchronized view contains the results of the analysis specifically requested by the user in the tree view. Both use the same statistical model (hypergeometric, binomial, χ^2 or Fisher's exact test, as appropriate).

Synchronized view at a custom abstraction level. However, the tree view by itself is insufficient. In this view, it is not possible to sort the results by GO terms, total number of genes or the *P*-value. This view is also inconvenient since in order to inspect all results, the user has to go through the entire tree manually looking for the significant biological processes, molecular functions and cellular components. In order to address this, another view, called synchronized view, has now been added to OE. This view is synchronized with the tree view. It displays the categories at the levels of abstraction chosen in the tree view as a bar graph and allows the user to sort them by GO terms, *P*-value or the total number of genes. When a category is collapsed, the number of genes for the category indicates the number of unique genes from the input list for the category itself as well as for its subcategories. In other words, when a subcategory is not visible, the genes annotated with the subcategory are considered as annotated with the current category. When a category is expanded, its subcategories are visible and corresponding bars are added in the synchronized view. When a category is expanded, if there are no genes from the input list annotated with the corresponding term, the number next to it becomes zero. In this situation, the bar corresponding to the term in the synchronized view is made invisible. Also, when a categories is collapsed, all the bars corresponding to its subcategory are made invisible in the synchronized view. OE also displays the results at the highest level of detail for all genes under the flat view.

Ability to accept and display fold changes. Users can now optionally specify the expression values for each gene obtained from their microarray experiment along with the list of genes. The gene ID and its expression value must be separated by a tab character. The expression values as specified by the user are displayed in the tree view, which greatly enhances the interpretation capabilities of the OE results. For example, instead of merely identifying apoptosis as a process that is significantly impacted, the user is now able to quickly understand the type of changes: if the apoptotic genes are mostly up-regulated, the apoptosis is stimulated, whereas if they are mostly down-regulated, the apoptosis is inhibited.

Supporting new input types. In addition to extending the support to 5 organisms and more than 172 commercial microarrays, OE is now able to support more types of input data. Previously, OE allowed users to submit a list of GenBank accession numbers, Affymetrix probe IDs or UniGene cluster IDs only. Now OE supports a list of gene names, gene symbols, Gene Ontology Consortium IDs and LocusLink IDs.

Links to the KEGG pathway database. It is crucial to understand the various biological pathways a gene is involved in, how the gene interacts with the other genes in the pathway and how it affects the expression of the interacting genes. The gene names in OE results are now hyper-linked to the KEGG pathway database, which graphically displays how the gene interacts with the other genes in a pathway and its effect on the other genes in the pathway.

Enhanced result files. Results files are now saved on the user's machine instead of being emailed to the user. The output files also contain much more details than before, including the GO ID of each term, the number of unique UniGene IDs for each GO term on the reference array, the LocusLink ID of each gene, the number of unique GenBank accession numbers for each GO term on the reference array and the official gene symbol. Users can now specify the type of information they want to store in the output files.

Additional features. There are many new features added for users' convenience. For example, the user can click on a single gene in the tree view and look up the other GO terms that the gene is annotated with. The user can also click on a GO term and look up a list of all the genes from the input list that are annotated with the GO term in a details window. UniGene cluster IDs and LocusLink IDs in the details window are hyper-linked to the UniGene database and the LocusLink database at NCBI. The details window now provides literature citations for each gene, with a brief summary. The citations are hyper-linked to the PubMed database. We have also addressed the feature requested by many users to be able to save the results as images, in addition to the tabular output files. The user can now save each of the graphs individually or select a set of terms and then save them as a GIF image.

Onto-Compare

Support for additional microarrays. All new microarrays added to the Onto-Tools database are also available in OC. The database now contains arrays for human, mouse, rat, *Drosophila* and *C.elegans*. The database now contains 172 focused arrays for the study of stress, oncogenes, angiogenesis, metastasis, prostate cancer, neurobiology, hematology, cytokines, apoptosis, cardiovascular systems and endocrine disruption.

Integrated GO browser. The results of OC are displayed in a GO tree as well as in a table format similar to the flat view in OE (see Figure 2). In the tree view, as a category is collapsed the genes annotated with its subcategories are considered to be annotated with it, and the total number of genes for the current category is increased appropriately. The results displayed under the table view consider each gene on the arrays as being associated only to the GO term it is annotated with.

Onto-Design

Search in the GO browser and support for keywords as input. In the previous release, OD required the user to submit the exact GO term relevant to the condition under study in

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 29, 2015

order to design an array. This requirement severely limited the usefulness of OD since it required the user to know exactly all GO terms of interest. The user was forced to go through the entire GO tree in order to find the GO terms of interest or search the GO Consortium website for the terms before using OD. In the new release of OD both limitations of OD are overcome. The user can now search the entire GO tree in order to look up the terms of interest. The user can also submit a list of keywords as an input file if the exact GO terms are not known. The keywords are case-insensitive. In such a case, OD returns the list of genes for all terms that contain the keywords. Along with keywords, one can also submit exact GO terms by selecting them in the GO tree or by simply adding them to the input file. The input file should be a simple ASCII text file with one keyword or one GO term per line.

Detailed result files. The output file format of OD is also modified to provide more details for each gene. OD results are now saved in two separate files. One of the output files contains the GO term and its corresponding list of genes. The other file contains the UniGene cluster ID, LocusLink ID, gene name and the reference accession number for each gene selected.

Tools

Implementation as standalone application. Each tool in the Onto-Tools ensemble is divided into two parts: server and client. The server part of the tools runs on a Jakarta Tomcat server. The server part of each tool is responsible for generating the results by accessing the Oracle 8*i* annotation database. The client part of each tool is responsible for displaying the results as well as browsing the results. The client part of Onto-Tools has now been implemented as a standalone application. This means it can run on the user's computer, unlike its previous release, which ran only as an applet in a Java-enabled web browser. The new Onto-Tools client is also enhanced using the Java Swing technology. This has greatly improved the performance and allowed the implementation of several new features. However, all Onto-Tools can still be run in a web browser. In this case, a security certificate is presented to the user. The user needs to accept the certificate in order to use the tools. This allows the application to save the results as regular files on the local machine and avoid the email transmission.

Improved design. One of the most important new features in the current release of Onto-Tools is the integration of the GO browser into OE, OC and OD. In order to integrate this GO browser, the client part of each tool needs the entire tree in a proper 'ready to display' format. This GO tree can be obtained from the server. However, the amount of data required to display the GO tree is enormous and the transfer of this data is very slow. In order to improve the performance of the tools, the client part of the tools requests the GO tree from the server as soon as the user logs in. In most cases, the time required for the user to navigate through the files on the client machine, select the input parameters and submit the request, as well as the time required by the server to process the request, is sufficient to pass the GO tree from the server to the client. However, in some cases it requires more time to pass the GO tree data from the server to the client than the time to submit the request by a user and to generate results by the server. In such cases, the Onto-Tools client will not be able to display the GO browser if the tree data is not available. In order to avoid 'race' conditions possibly causing failures in displaying the results, we employed various synchronization techniques that ensure that the Onto-Tools do not attempt to display results until the GO tree data become available on the client side. Note that the GO tree is required to be passed to the client only once and will be re-used for subsequent requests.

Concurrent usage of tools. Another major enhancement to the features of Onto-Tools is the concurrent use of tools. In the previous release, the user could use only one tool at a time. In addition, if the user ran the same tool (e.g. OE) twice with two different sets of genes, it was not possible to compare the results of the two requests. The new release of Onto-Tools not only allows users to run more than one tool concurrently, but also allows them to submit more than one concurrent request for the same tool.

DISCUSSION

A potential pitfall of our ontological analysis approach is that ontology can be biased in favor of certain genes or pathways. Indeed, the number of annotations available is directly proportional to the number of experiments performed with those genes or pathways. Some biological processes are more intensively studied, thus generating more data (e.g. recently apoptosis has been a much-studied process). If more data about a specific process are available, this process is more likely to appear in the results of OE since more genes are known to be associated with it. This bias can be eliminated by using the recently added statistical analysis (14,25). The purpose of this analysis is precisely to interpret the results in the light of the amount of known information about each functional category. Note that a biological category with a low *P*-value (non-random) is, in most cases, a biologically important process. However, in some cases, a low *P*-value may also reflect a non-random, non-biological process somewhere in the data-processing pipeline (e.g. an incorrect normalization method used to select the differentially regulated genes provided as input to OE).

Another potential pitfall is that the results of any analysis are limited by the availability of accurate annotations. In many experiments involving less studied species, it is possible for OE to return no useful results simply because the functions of the genes provided as input are not known. It is acknowledged that the existing data are incomplete. However, it is worth noting that in spite of the limitations of the ontological analysis approach offered by Onto-Tools, this is undeniably better than the only other alternative, which is the manual retrieval of annotation on a gene-by-gene basis from several databases.

Finally, our ontological analysis approach can be criticized because certain genes are more important than others, so the sheer number of genes may not tell the whole story. In certain situations, a small change in the expression of a single gene (e.g. a transcription factor) may trigger large ripple effects. It is clear that a quantitative analysis of genes, while very informative, may not always be able to capture the entire complexity of certain cellular processes. Overall, we must emphasize that there is no substitute for human knowledge and intelligence. Once one has the results of any of the Onto-Tools, one has to analyze them carefully from a biological perspective. A computer analysis, no matter how sophisticated, will never

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 29, 2015

capture the entire complexity of a living organism, and a computer, no matter how expensive, will never be a substitute for the inquisitive and analytic mind of a trained researcher.

Since the public release of OE in 2001 (13), other programs with functionality similar to OE have been made available. GoMiner, proposed in 2002 as a GO browser, recently added Fisher's exact test to calculate statistically significant GO categories (26). In addition to Fisher's exact test, OE provides hypergeometric, binomial and χ^2 tests and allows the user to switch between them as needed (14). This is important since it has been shown that Fisher's exact test is not optimal and should be used only when very few genes are involved (27). The same group recently provided a separate tool, MatchMiner, which offers functionality similar to that of OT. MatchMiner converts accession numbers, UniGene cluster IDs and LocusLink IDs into HUGO gene names, which in turn can be used as input to GoMiner. In order to create a functional profile for a list of accession IDs in GoMiner, the user needs to run two separate programs. In contrast, OT's integration allows one-click access to the same functionality. An interesting feature of GoMiner is that if the input human gene(s) do not have annotations, GoMiner returns results based on mouse annotations.

Released in 2003, DAVID and EASE are two other tools that perform an ontological analysis similar to that originally proposed by OE (28,29). DAVID reports results based on the input type; hence, two accession numbers mapped to the same UniGene cluster will be counted twice. Two other recent tools are Vlad (http://www.informatics.jax.org/~jer/vlad/) and GO::Term Finder (http://search.cpan.org/dist/GO-TermFinder/lib/GO/TermFinder.pm). Vlad uses a fixed depth in GO but provides a graphical view of the GO subgraph involving the input genes. GO::TermFinder analyzes only one category at a time (i.e. biological process or molecular function or cellular component, but not all). Vlad, TermFinder and GoMiner (26) do not accept accession numbers, UniGene cluster IDs or LocusLink IDs as input. None of the tools above has our chromosome view, customized abstraction level or pointers to relevant publications, for example. The functionalities of the newer Onto-Tools, Compare, Design and Miner, are not yet available elsewhere.

ACKNOWLEDGEMENTS

This work has been supported by the following grants: NSF DBI-0234806, DOD DAMD 17-03-02-0035, NIH(NCRR) 1S10 RR017857-01, MLSC MEDC-538 and MEDC GR-352, NIH 1R21 CA10074001, 1R21 EB00990-01 and 1R01 NS045207-01. Onto-Tools currently runs on equipment provided by Sun Microsystems under the grant EDU 7824-02344-US.

REFERENCES

1. Bankier,A.T. (2001) Shotgun DNA sequencing. *Methods Mol. Biol.*, **167**, 89–100.
2. Craig,J.C., Adams,M.D., Myers,E.W., Li,P.W., Mural,R.J., Sutton,G.G., Smith,H.O., Yandell,M., Evans,C.A., Holt,R.A., *et al.* (2001) The sequence of the human genome. *Science*, **291**, 1304–1351.
3. Eisen,M.B., Spellman,P.T., Brown,P.O. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci., USA*, **95**, 14863–14868.
4. Golub,T.R., Slonim,D.K., Tamayo,P., Huard,C., Gaasenbeek,M., Mesirov,J.P., Coller,H., Loh,M.L., Downing,J.R., Caligiuri,M.A., Bloomfield,C.D. and Lander,E.S. (1999) Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, 531–537.
5. Lockhart,D.J., Dong,H., Byrne,M.C., Folletie,M.T., Gallo,M.V., Chee,M.S., Mittmann,M., Want,C., Kobayashi,M., Horton,H. and Brown,E.L. (1996) DNA expression monitoring by hybridization of high density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.
6. Schena,M., Shalon,D., Davis,R. and Brown,P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.
7. Schena,M. (2000) *Microarray Biochip Technology*. Eaton Publishing, Westborough, MA.
8. Shalon,D., Smith,S.J. and Brown,P.O. (1996) A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.*, **6**, 639–645.
9. Tamayo,P., Slonim,D., Mesirov,J., Zhu,Q., Kitareewan,S., Dmitrovsky,E., Lander,E.S. and Golub,T.R. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci., USA*, **96**, 2907–2912.
10. Eisenberg,D., Marcotte,E.M., Xenarios,L. and Yeates,T.O. (2000) Protein function in the post-genomic era. *Nature*, **405**, 823–826.
11. Lockhart,D.J. and Winzeler,E.A. (2000) Genomics, gene expression and DNA arrays. *Nature*, **405**, 827–836.
12. Vukmirovic,O.G. and Tilghman,S.M. (2000) Exploring genome space. *Nature*, **405**, 820–822.
13. Khatri,P., Draghici,S., Ostermeier,G.C. and Krawetz,S.A. (2002) Profiling gene express with Onto-Express. *Genomics*, **79**, 266–270.
14. Draghici,S., Khatri,P., Martins,R.P., Ostermeier,G.C. and Krawetz,S.A. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.
15. Draghici,S., Khatri,P., Shah,A. and Tainsky,M.A. (2003) Assessing the functional bias of commercial microarrays using the Onto-Compare database. *BioTechniques* (suppl.), 55–61.
16. Draghici,S., Khatri,P., Bhavsar,P., Shah,A., Krawetz,S.A. and Tainsky,M.A. (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, **31**, 3775–3781.
17. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. and Eppig,J.T., *et al.* (2000) Gene Ontology: tool for the unification of biology. *Nature Genet.*, **25**, 25–29.
18. BD PowerBlot Western Array Screening Service (2002) *Technical Report*. BD Biosciences, Franklin Lakes, NJ.
19. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
20. Kanehisa,M., Goto,S., Kawashima,S. and Nakaya,A. (2002) The KEGG databases at GenomeNet. *Nucleic Acids Res.*, **30**, 42–46.
21. Ogata,H., Goto,S., Sato,K., Fujibuchi,W., Bono,H. and Kanehisa,M. (1999) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **27**, 29–34.
22. BioCarta (2004) Charting pathways of life. *Technical Report*. BioCarta, San Diego, CA.
23. Krieger,C.J., Zhang,P., Mueller,L.A., Wang,A., Paley,S., Arnaud,M., Pick,J., Rhee,S.Y. and Karp,P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **32**, D438–D442.
24. Apweiler,R., Bairoch,A., Wu,C.H., Barker,W.C., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. and Magrane,M. *et al.* (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.*, **32**, D115–D119.
25. Draghici,S. (2003) *Data Analysis Tools for Microarrays*. Chapman and Hall/CRC Press, London, UK.
26. Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C. and Lababidi,S. *et al.* (2003) A resource for biological interpretation of genomic and proteomic data. *Genome Biol.*, **4**, R28.
27. Man,M.Z., Wang,Z. and Wang,Y. (2000) POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics*, **16**, 953–959.
28. Hosack,D.A., Dennis,G., Sherman,B.T., Lane,H.C. and Lempicki,R.A. (2003) Identifying Biological Themes within Lists of Genes with EASE. *Genome Biol.*, **4**, P4.
29. Dennis,G., Sherman,B.T., Hosack,D.A., Yang,J., Gao,W., Lane,H.C. and Lempicki,R.A. (2003) DAVID: database for annotation, visualization, and integrated discovery. *Genome Biol.*, **4**, P3.

Downloaded from http://nar.oxfordjournals.org/ at Michigan State University on November 29, 2015